Evaluation of Meteorite Amino Acid Analysis Data using Multivariate Techniques

Gene D McDonald, Michael C. Storrie-Lombardi, and Kenneth H. Nealson

Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, CA 91109

for submission to *Meteoritics and Planetary Science*

Abstract

The amino acid distributions in the Murchison carbonaceous chondrite, Mars meteorite ALH84001, and ice from the Allan Hills region of Antarctica are shown, using a multivariate technique known as Principal Component Analysis (PCA), to be statistically distinct (2.1-3.6σ) from the average amino acid compositions of 101 terrestrial protein superfamilies. The Allan Hills ice amino acid distribution clusters with those of ALH84001 samples, and is distinct (3.5-3.8σ) from the proteins. This indicates that contaminant biological amino acids in Antarctic ice have undergone fractionation in the geochemical environment. The several Murchison analyses examined have intracluster distances less than 1σ, suggesting that terrestrial contamination in these samples is minimal. PCA accomplishes these discriminations using either raw amino acid distribution data or amino acid elemental composition. This work demonstrates that PCA and multivariate analysis in general can be useful for the interpretation of possible biosignatures in extraterrestrial samples.

Introduction

Over 70 different amino acids, including at least six of the twenty protein amino acids, have been identified in extracts of the Murchison carbonaceous chondrite (Cronin and Pizzarello 1983; Cronin et al. 1988). The non-protein amino acids in Murchison are accepted as extraterrestrial material, even in light of recent reports of small enantiomeric excesses of some of the non-protein amino acids (Cronin and Pizzarello 1997). Aside from some initial uncertainties due to laboratory contamination, the protein amino acids in Murchison are also believed to be extraterrestrial, based on both their racemic chiral signatures and their distinctly non-terrestrial carbon and nitrogen isotope compositions (Cronin et al. 1988). There have been reports claiming large enantiomeric excesses of some protein amino acids in Murchison (Engel and Nagy, 1982; Engel et al. 1990; Engel and Macko, 1997), but these claims remain controversial (Bada, et al. 1983; Pizzarello and Cronin, 1998).

This paper introduces principal component analysis (PCA) as an approach to the analysis of meteoritic amino acid distributions. The PCA profiles of the amino acid compositions of terrestrial biological protein families are compared with those of the suite of amino acids detected in the Murchison carbonaceous chondrite. PCA comparison is also used to examine the amino acid composition of the Mars meteorite ALH84001 (Bada et al. 1998). The presence of organic compounds (polycyclic aromatic hydrocarbons rather than amino acids) was used by McKay et al. (1996) to argue that ALH84001 contained evidence of fossil life on Mars. Bada et al. (1998), however, examined the amino acid content of the meteorite and interpreted their data as indications of contamination by ice-borne terrestrial contaminants. This conclusion was based on both the abundance profile and the D/L ratios of the amino acids in the meteorite. Detection of

^{14}C in the organic material from ALH84001 (Jull et al. 1998) further supports the contention that most if not all of the organic content of the meteorite is biologically derived material from Earth.

Methods

To illustrate the application of the PCA method to extraterrestrial organic analysis, comparisons were made of the amino acid distributions found in terrestrial proteins with the distributions of amino acids found in the Murchison carbonaceous chondrite, the Mars meteorite ALH84001, and one sample of ice from the Allan Hills region of Antarctica. The average amino acid compositions of 101 protein superfamilies were taken from the compilations of Dayhoff (1972, 1978). The compositions of protein amino acids in the Murchison carbonaceous chondrite were taken from several published analyses. From Cronin and Pizzarello (1983) came a consensus protein amino acid composition for Murchison, based on work published up to that time, as well as side-by-side analyses by Cronin and Pizzarello (1983) of three Murchison samples, labelled "Field Museum", "ASU", and "Smithsonian". Murchison protein amino acid compositions reported by Engel and Nagy (1982), Engel et al. (1990), and Engel and Macko (1997) were also used. Amino acid compositions of ALH84001 and Allan Hills ice were taken from Bada et al. (1998).

In this paper a classical Multivariate Analysis known as *Principal Component Analysis* (PCA) is used. Also known as the Karhunen-Loéve or Hotelling transform, PCA identifies linear combinations of raw parameters accounting for maximum variance in the data set. The algorithm is easily implemented with a variety of free packages available in C or Fortran over the Internet as well as being included in commercial statistical packages

such as Statview (Abacus Concepts), the software employed in this paper. PCA also makes it possible to reduce the total number of variables by creating a minimum set of factors (combinations of the original variables) to account for the maximum amount of information in a data set. The technique first determines the Pearson correlation matrix for the data and then calculates the eigenvector and eigenvalues for the matrix.

To illustrate, consider the simplest 2-variable case in geometrical terms. PCA performs a linear least-squares fit of a projection through the original data cloud created on a 2-dimensional x-y plot. Some measure, usually the total Euclidean distance between each data point and the nearest point on the projection, serves as an estimate of the amount of variance accounted for by this first projection or axis. In matrix algebra nomenclature these are the first eigenvector and its associated eigenvalue. If significant variance remains, PCA repeats the process with a new projection orthogonal to the first. The new x and y values are obtained from the distance along the new axis (x') to its intersection with an orthogonal projection to each data point and the distance from the projection to the associated data (y').

In the work presented in this paper, the input vector of raw parameters to the PCA algorithm was either the amino acid frequencies, or the hydrogen (H)/carbon (C)/nitrogen (N) compositions, of the amino acids found in the meteorite, protein, and ice samples. In the latter case, the H/C and N/C ratios were also included as part of the input vector. In either case, denoting any input vector as $S_{i,l}$, where i represents n input parameters for each vector and l indexes the m training vectors, PCA first finds the mean spectrum, $\Delta S_{i,l}$, calculates the difference between this mean and each individual vector, estimates the parameter covariance matrix, $C_{j,k}$, and decomposes this into constituent eigenvalues and normalized eigenvectors. If the decomposition is restricted to the first p terms, the method yields the optimum (in a least squares sense) linear reconstruction of the input signal using the fewest, p, parameters.

In this study, it was possible to obtain separation of the protein and meteorite/ice samples using as few as three PCA factors. A key advantage of PCA is its ability to transform a wide variety of input vectors to a common statistical metric output (Storrie-Lombardi et al., 1994). Specifically, data sets shift to zero mean and unit variance. This transformation makes it possible for us to assess the statistical significance of any apparent clustering in a multi-dimensional (in our case 3-D) data representation. In addition, since the covariance matrix is an average over many vectors and since noise is uncorrelated between vectors, the method is quite robust to modest amounts of noise.

Results and Discussion

In the first round of PCA analyses, the raw frequency distributions for the 20 terrestrial protein amino acids in each sample were used.

-------Figure 1 about here--------

Figure 1A shows the entire data set plotted using values for PCA factors 1, 2, and 5. As shown by the weighting values in Table I, factor 1 is primarily composed of interactions between glutamine, asparagine and tyrosine. Factor 2 represents methionine, tryptophan, and to a lesser extent phenylalanine and tyrosine, while factor 5 has a strong negative correlation with serine and a positive correlation with glutamic acid. Using simple probability distributions, PCA factor extraction separated the meteorite and ice samples from the protein families relatively well. As shown in Table II, the intercluster distances between the protein cluster (PC) and the two meteorite clusters (MC and AC) are 2.2-3.2 sigma.

Figure 1B shows only the region of parameter space containing the meteorite samples. The separation between meteorites and protein superfamilies is almost complete, with only two outlying protein data points separated from the protein cluster by the Murchison cluster and the ALH cluster. The intercluster distance between Murchison and ALH84001 is 4.5σ (Table II), a very significant separation. There is some elongation within each cluster along the F5 axis, resulting in better actual separation between the Murchison sample cluster and the ALH84001/Allan Hills ice cluster than the intercluster distances suggest.

While use of the amino acid distributions can provide us with significant information, the technical challenges of adapting analytical methods to spacecraft use may in some cases prevent acquisition of molecular abundance data from extraterrestrial samples. Elemental compositions

-------Figure 2 about here--------

of samples, on the other hand, may be easier to obtain in an *in situ* analysis. Even for a returned sample, elemental analysis may be easier to accomplish due to small sample sizes. Figure 2 shows the result of reducing the amino acid data sets to abundances of hydrogen (H), carbon (C), and nitrogen (N), plus the ratios H/C and N/C. As shown in Table I, factor 1 represents a positive contribution from H and C, and a negative contribution from H/C. Factor 2 represents N and N/C, while factor 3 represents contributions from H and H/C. In Figure 2A the separation between the meteorite samples and the protein superfamilies is more distinct than that given by the unreduced amino acid abundance data (Fig. 1). The distance between the borders of the protein and Murchison clusters in Fig. 2

corresponds to 1.7-1.8 standard deviations. As seen in Figure 2B, the separation between Murchison and ALH samples is not quite as large, but definite clusters can still be discerned.

As mentioned above, Engel and Nagy (1982) reported large excesses of L-alanine in their Murchison analysis, a claim that has been both supported (Engel et al. 1990, Engel and Macko 1997) and disputed (Bada et al. 1983, Pizzarello and Cronin 1998). In the analyses presented in this paper, the Murchison samples cluster rather tightly, with intracluster distances of 0.38 and 0.65 and distances from the protein cluster of 2.2-2.6σ. It appears from these analyses that terrestrial contamination of these Murchison samples is minimal.

The Allan Hills ice sample groups tightly with the ALH84001 samples in both analyses, supporting the hypothesis that the amino acids detected in ALH84001 and those present in Antarctic ice have the same source, most likely terrestrial biological contamination (Bada et al. 1998). However, the ice sample is also consistently distinct from the protein amino acid clusters, with distances of greater than 3σ in both analyses. There are two possible explanations for the separation between the ALH84001/Antarctic ice cluster and the protein cluster. The 84001/ice amino acid distribution may be the result of terrestrial biological contamination, with a significant amount of geochemical fractionation of amino acids taking place at some point in the process of airborne or seaborne transport from lower latitudes and deposition onto Antarctic ice. Terrestrial amino acids deposited in the Antarctic would probably have long atmospheric or oceanic residence times, allowing opportunities for selective destruction or rainout of amino acids to occur.

Alternately, the observed 8400l/ice distribution could be the result of input of chondritic material to the ice in the form of micrometeorites (Brinton et al. 1998). If this explanation is correct, the amino acid distribution of the ice should closely resemble the Murchison distribution. In the distribution analysis (Fig. 1 and Table II) the distance from the ice sample to the protein cluster is less by more than 1σ than the distance from the ice to the Murchison cluster. In the elemental analysis, however (Fig. 2 and Table III), the ice is closer to the Murchison cluster by 1.5σ. Thus the provenance of the background amino acids in Antarctic ice appears to be complex, with perhaps multiple sources.

Implications for Future Mars Sample Analysis

The analysis of extraterrestrial samples for organic signatures of past or present life presents several problems. One of these is deciding which organic compounds to target. No sample is ever available in unlimited quantities, whether analyzed *in situ* or returned to Earth, and the sample extraction and preparation techniques used to isolate one class of organic compounds often render the sample unusable for other analyses. Another challenge is distinguishing bona fide extraterrestrial organic material from terrestrial contamination, either carried on a spacecraft or present in the terrestrial environment to which the sample is exposed. This problem becomes most acute in the case of meteorites, which may have been exposed to the environment and biota of Earth for periods of up to 10^4 years. A final problem, once extraterrestrial organics have been identified, is in separating biologically-derived molecules from those produced by abiotic syntheses in the interstellar medium, on meteorite parent bodies, or in planetary atmospheres and oceans. Stable isotope signatures have been used in some cases to help overcome these last two

problems, but current limitations in sensitivity and in miniaturization of instruments for in situ analysis restrict the use of isotope techniques.

Principal Component Analysis and other multivariate analysis techniques show great potential in discrimination of biological from non-biological material in extraterrestrial samples. The strength of this approach is the ability to use molecular distributions and/or elemental composition data. Profiles of different biomarker compounds, e.g. amino acids, fatty acids, etc., can be combined directly or by reduction to elemental compositions. In circumstances where identification of individual compounds is not feasible, elemental analysis data can still be used to construct biomarker profiles. The technique also can potentially combine chromatographic and spectroscopic data from a single sample into a unified data set, resulting in a true multidimensional biosignature profile.

PCA of amino acid compositions from Murchison and ALH84001 clearly produces two linearly separable clusters. One cluster is populated by the Murchison samples, while the second encompasses the ALH84001 samples as well as the single Antarctic ice sample. Since the separation metric measures cluster differences in terms of data variance, graphically represented as sigma values, the technique appears capable of providing a quantitative method for assessing terrestrial contamination of extraterrestrial samples.

Acknowledgements

This work was supported by the NASA Exobiology Program and by the Jet Propulsion Laboratory. The authors thank A. Tsapin for helpful comments.

References

Bada J. L., Cronin J. R., Ho M.-S., Kvenvolden K. A., Lawless J. G., Miller S. L., Oro J. and Steinberg S. (1983) On the reported optical activity of amino acids in the Murchison meteorite. *Nature* **301**, 494-497.

Bada J. L., Glavin D. P., McDonald G. D. and Becker L. (1998). A search for endogenous amino acids in martian meteorite ALH84001. *Science* **279**, 362-365.

Brinton K. L. F., Engrand C., Glavin D. P., Bada J. L., and Maurette M. (1998). A search for extraterrestrial amino acids in carbonaceous Antarctic micrometeorites. *Origins Life Evol. Biosphere* **28**, 413-424

Cronin J. R. and Pizzarello S. (1983) Amino acids in meteorites. *Adv. Space Res.* **3(9)**, 5-18.

Cronin J. R., Pizzarello S. and Cruikshank D. P. (1988). Organic matter in carbonaceous chondrites, planetary satellites, asteroids and comets. In *Meteorites and the Early Solar System* (eds. J. F Kerridge and M. S. Matthews), pp. 819-857. Univ. Ariz. Press, Tucson.

Cronin J. R. and Pizzarello S. (1997) Enantiomeric excesses in meteoritic amino acids. *Science* **275**, 951-955.

Dayhoff M. O. (1972). *Atlas of Protein Sequence and Structure*, v. 5, Natl. Biomed. Res. Found.

Dayhoff M. O. (1978). *Atlas of Protein Sequence and Structure*, v. 5, supl. 3, Natl. Biomed. Res. Found.

Engel M. H. and Nagy B. (1982) Distribution and enantiomeric composition of amino acids in the Murchison meteorite. *Nature* **296**, 837-840.

Engel M. H., Macko S. A., and Silfer J. A. (1990) Carbon isotope composition of individual amino acids in the Murchison meteorite. *Nature* **348**, 47-49.

Engel M. H. and Macko S. A. (1997) Isotopic evidence for extraterrestrial non-racemic amino acids in the Murchison meteorite. *Nature* **389**, 265-268.

Jull A. J. T., Courtney C., Jeffrey D. A., and Beck J. W. (1998) Isotopic evidence for a terrestrial source of organic compounds found in Martian meteorites Allan Hills 84001 and Elephant Moraine 79001. *Science* **279**, 366-369.

McKay D. S., Gibson E. K. Jr., Thomas-Keprta K. L., Vali H., Romanek C. S., Clemett S. J., Chiller X. D. F., Maechling C. R., and Zare R. N. (1996) Search for past life on Mars: Possible relic biogenic activity in Martian meteorite ALH84001. *Science* **273**, 924

Pizzarello S. and Cronin J. R. (1998) Alanine enantiomers in the Murchison meteorite. *Nature* **394**, 236.

Storrie-Lombardi M. C., Irwin M., von Hippel T., and Storrie-Lombardi L. J. (1994) Spectral classification with principal component analysis and artificial neural networks. *Vistas in Astronomy* **38**, 331-340.

Table I

Oblique Solution Reference Structures

Amino acid distribution

	F1	F2	F5
ALA	-.217	-.411	.164
ARG	-.090	-.086	-.010
ASN	.786	-.030	-.083
ASP	-.01	-.01	.180
CYS	.041	-.169	.167
GLN	.931	-.031	.059
GLU	-.180	.028	.708
GLY	-.243	-.239	-.051
HIS	-.014	.104	-.080
ILE	.029	-.200	.042
LEU	.165	-.129	.072
LYS	.093	-.048	.037
MET	.010	.839	-.129
PHE	-.051	.539	.117
PRO	-.211	.305	.305
SER	-.082	.136	-.934
THR	-.01	-.01	-.067
TRP	.077	.807	-.044
TYR	.573	.454	-.050
VAL	-.124	.017	.062

Amino acid H/C/N compositions

	F1	F2	F3
H	.928	.268	.259
C	.995	.072	.024
N	.508	.856	-.087
H/C	-.885	.304	.352
N/C	-.659	.738	-.140

Table II

Cluster distances in sigma for raw amino acid distributions

Intercluster

	MC	AC	PC	Ice
MC	0			
AC	4.69	0		
PC	2.18	3.03	0	
Ice	5.34	0.66	3.67	0

Intracluster

MC	0.52 ± 0.22
AC	1.51 ± 0.10
PC	1.37 ± 1.06

PC - protein cluster; AC - ALH84001 cluster; MC - Murchison cluster

Table III

Cluster distances in sigma for amino acid elemental compositions

Intercluster

	MC	AC	PC	Ice
MC	0			
AC	1.50	0		
PC	2.62	3.60	0	
Ice	1.54	0.27	3.47	0

Intracluster

MC	0.65 ± 0.29
AC	0.51 ± 0.03
PC	1.12 ± 1.01

PC - protein cluster; AC - ALH84001 cluster; MC - Murchison cluster

Figure captions

Figure 1. PCA analysis of Murchison, ALH84001, protein, and Antarctic ice amino acid compositions using raw distribution data. A) Parameter space including all data points, B) Close-up of parameter space containing meteorite clusters.

Figure 2. PCA analysis of Murchison, ALH84001, protein, and Antarctic ice amino acid compositions using amino acid data reduced to H/C/N ratios. A) Parameter space including all data points, B) Close-up of parameter space containing meteorite clusters.

Figure 1A
PCA analysis of raw amino acid distribution

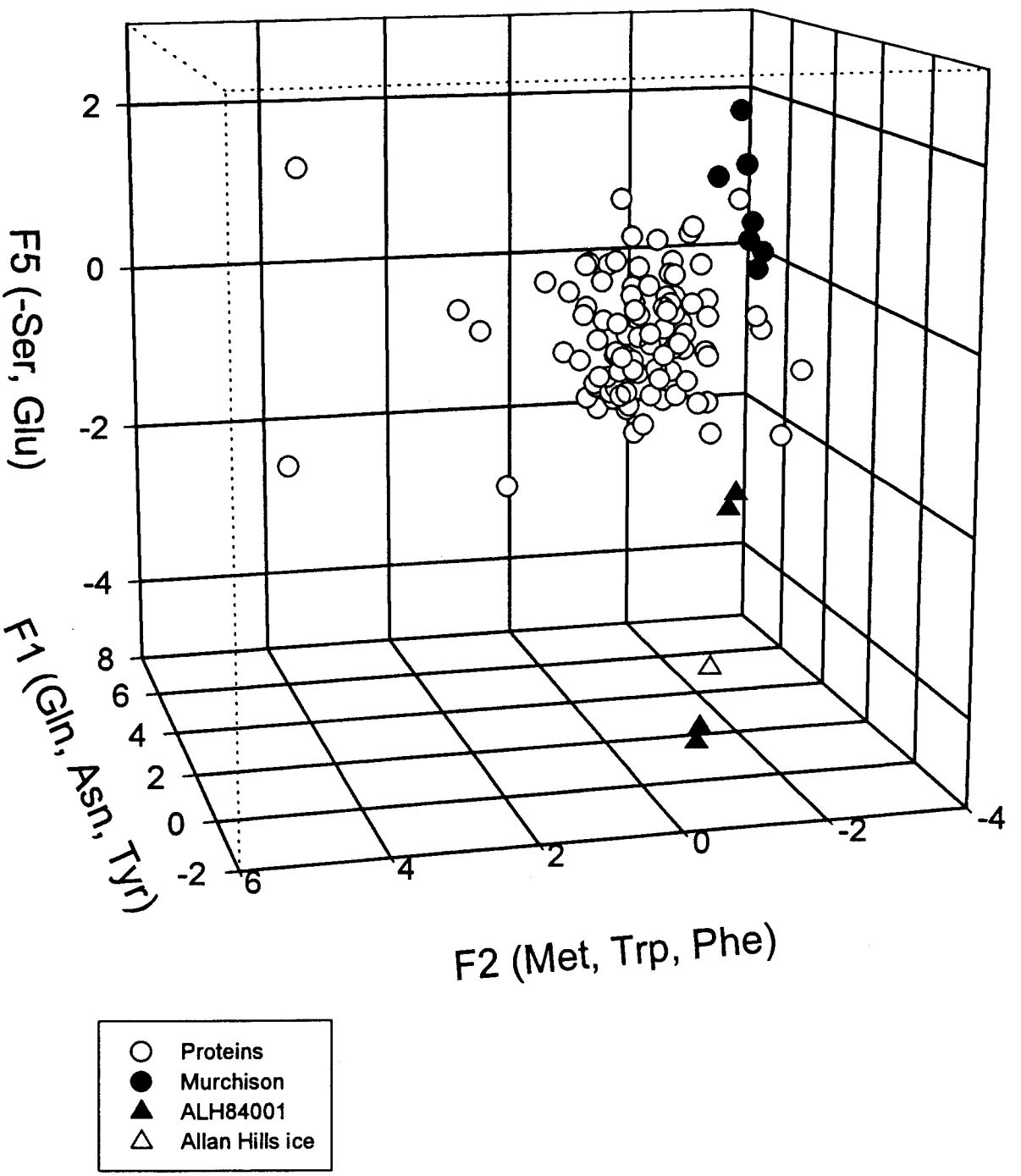

Figure 1B
PCA analysis of raw amino acid distribution

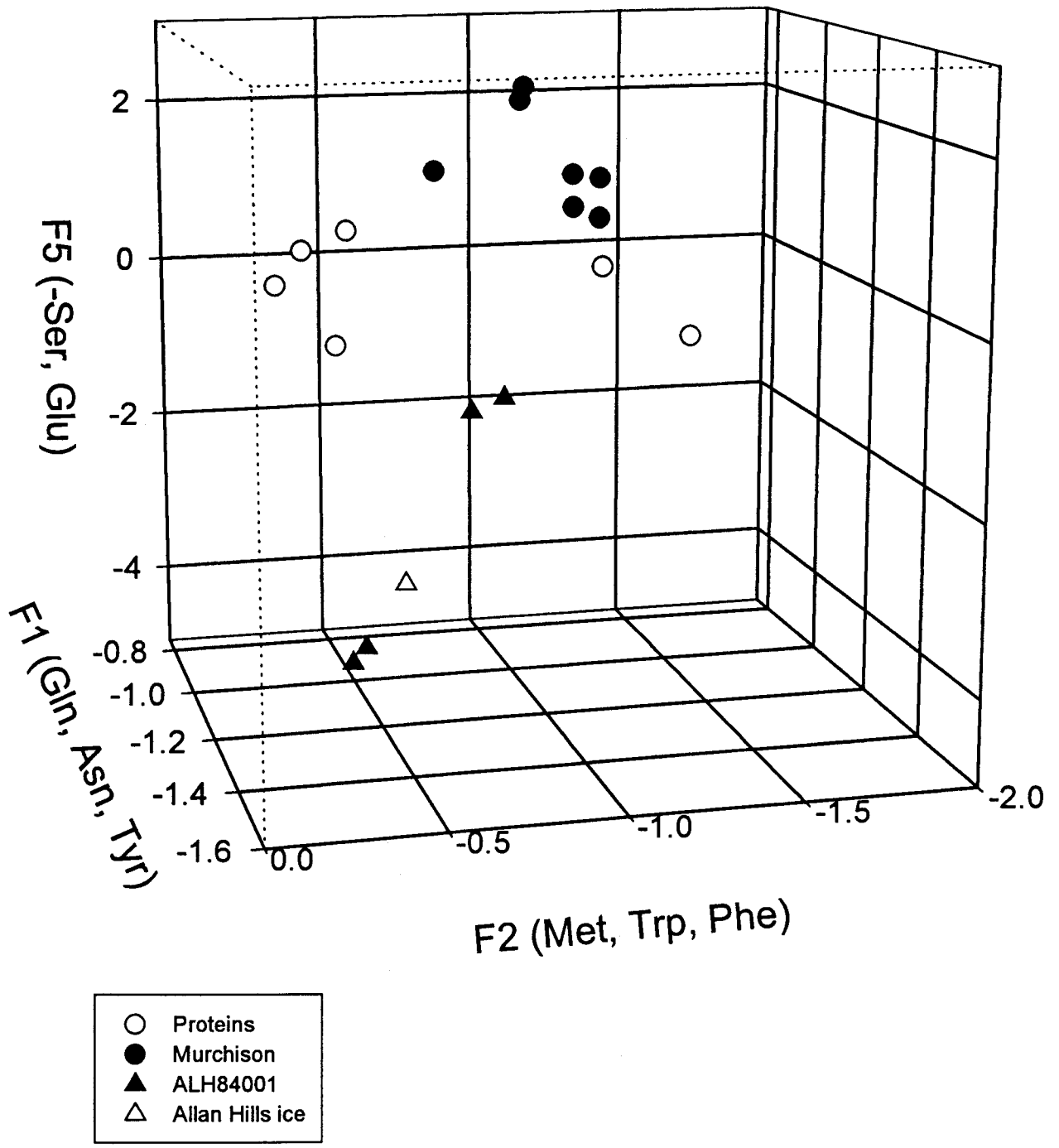

Figure 2A
PCA analysis of amino acid H/C/N composition

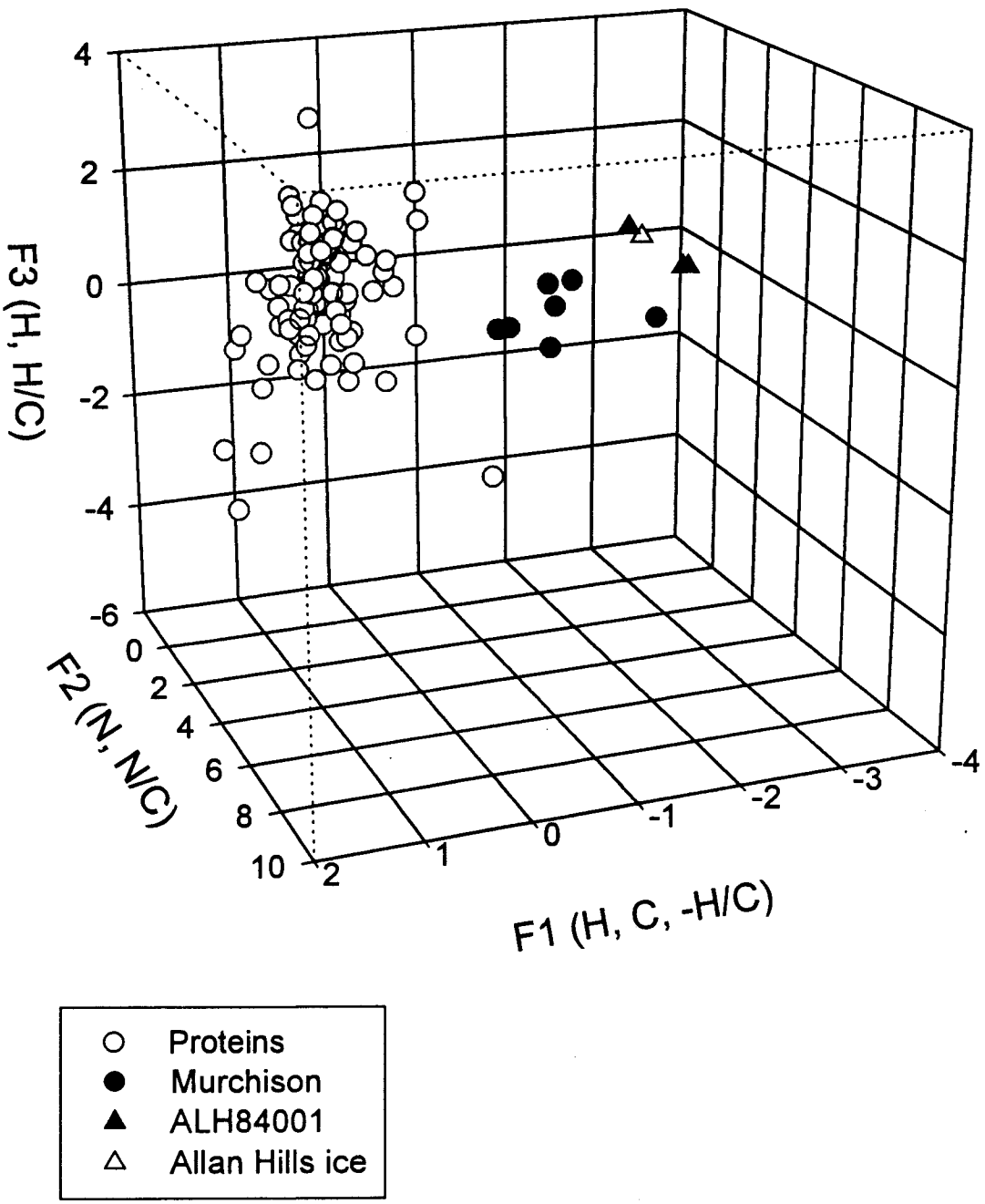

Figure 2B
PCA analysis of amino acid H/C/N composition